Does error control suppress spuriosity?

Aves, Mark A.; Griffiths, David; Higham, Desmond. J.

Published in:
SIAM Journal on Numerical Analysis

DOI:
10.1137/S0036142994276980

Publication date:
1997

Does Error Control Suppress Spuriosity?

Mark A. Aves * David F. Griffiths * Desmond J. Higham *

Abstract

In the numerical solution of initial value ordinary differential equations, to what extent does *local* error control confer *global* properties? This work concentrates on global steady states, or fixed points. It is shown that for systems of equations, *spurious* fixed points generally cease to exist when local error control is used. For scalar problems, on the other hand, locally adaptive algorithms generally avoid spurious fixed points by an indirect method—the stepsize selection process causes spurious fixed points to be unstable. However, problem classes exist where, for arbitrarily small tolerances, stable, spurious fixed points persist, with significant basins of attraction. A technique is derived for generating such examples.

1 Introduction

Numerical analysts approximate continuous flows by discrete maps. When the approximations are computed over long time intervals, it is well known that the discrete map can converge to a *spurious* steady state—a solution that is unrelated to the continuous problem. Such a state of affairs is clearly best avoided, and a great deal of attention has been paid recently to this phenomenon.

Several authors have considered the case of autonomous, initial value, ordinary differential equations (ODEs)

$$y'(t) = f(y(t)), \quad t > 0, \quad y(0) = y_0 \in \mathbb{R}^m. \tag{1.1}$$

Applying an s-stage explicit Runge-Kutta (ERK) formula with constant stepsize h to this problem produces approximations $y_n \approx y(t_n)$, with $t_n = nh$, satisfying

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \tag{1.2}$$

where

$$\begin{aligned} k_1 &= f(y_n), \\ k_i &= f(y_n + h \sum_{j=1}^{i-1} a_{ij} k_j), \quad 2 \le j \le s. \end{aligned}$$

*Department of Mathematics and Computer Science, University of Dundee, Dundee, DD1 4HN, Scotland. Email addresses: maves@mcs.dund.ac.uk, dfg@mcs.dund.ac.uk, dhigham@mcs.dund.ac.uk.

Here the coefficients $\{b_i, a_{ij}\}$ define a particular formula. The relation (1.2) can be regarded as a one-step map

$$y_{n+1} = y_n + h\Phi(h, y_n), \tag{1.3}$$

where the increment function Φ depends upon f. If $\Phi(h^*, y^*) = 0$ but $f(y^*) \neq 0$, then y^* is a spurious fixed point (of period one).

If a spurious fixed point is linearly stable, then, for a certain range of initial values y_0, the numerical solution will be attracted to this spurious value. Many examples of such spurious behaviour have been constructed ([1, 6, 7, 13]) and it has been found that "spuriousity" can occur even when the stepsize is chosen to satisfy the constraints imposed by absolute stability analysis. Related work by Hairer et al. [2] also looked at the Runge-Kutta process's propensity for spurious behaviour. In particular, these authors showed that any explicit Runge-Kutta formula other than forward Euler can produce a spurious fixed point.

Although this potential for spurious solutions is worrying, it should be noted that computing the residual, $f(y^*)$, gives a simple a posteriori check on the validity of a constant steady state. Further, Humphries [6] has shown that, under mild assumptions about f, any fixed point that exists for arbitrarily small stepsizes must become unbounded as $h \to 0$. Hence, repeating the integration with a smaller stepsize will ultimately make spurious behaviour apparent.

The work mentioned above deals with the behaviour of constant stepsize algorithms, and is relevant to many applications in science and engineering, particularly in the solution of semi-discretised, nonlinear, partial differential equations. However, as several authors have noted, standard software for ODEs does not use constant stepsizes. Instead, the stepsize $h_n := t_{n+1} - t_n$ is varied according to a local error criterion. In the case of explicit Runge-Kutta methods, the main formula (1.2) is coupled with a secondary formula, to give

$$\widehat{y}_{n+1} = y_n + h_n \sum_{i=1}^{s} \widehat{b}_i k_i. \tag{1.4}$$

Here $\widehat{y}_{n+1}$ is the result of a different ERK formula applied at y_n. (The order of the secondary formula may be higher or lower than that of the main formula.) We may write (1.4) in a manner analogous to (1.3):

$$\widehat{y}_{n+1} = y_n + h_n \Psi(h_n, y_n). \tag{1.5}$$

An error estimate for the step is given by either

$$\mathrm{est}_{n+1} = || \, y_{n+1} - \widehat{y}_{n+1} \, ||, \tag{1.6}$$

or

$$\mathrm{est}_{n+1} = || \, y_{n+1} - \widehat{y}_{n+1} \, ||/h_n. \tag{1.7}$$

The error estimate (1.6) is referred to as an error-per-step (EPS) estimate, while (1.7) is called an error-per-unit-step (EPUS) estimate.

The error estimate is used for two purposes—error control and stepsize selection, and almost all software employs the same basic strategy. If est_{n+1} satisfies the criterion $\mathrm{est}_{n+1} \leq \tau$, where τ is a user-supplied tolerance parameter, then the step is accepted.

Otherwise the step is rejected and re-computed with a smaller stepsize (until the condition $\text{est}_{n+1} \leq \tau$ becomes true). The usual formula for the next stepsize is

$$h_{n+1} = \theta \left(\frac{\tau}{\text{est}_{n+1}} \right)^{1/q} h_n. \tag{1.8}$$

Here q is an integer that is determined from the Runge-Kutta formulas. (It is the largest integer such that $\text{est}_{n+1} = O(h_n^q)$.) The constant safety factor $\theta \in (0,1)$ is included in an attempt to avoid rejecting too many steps. Values of θ between .8 and .9 are typical. The formula (1.8) can be justified by an asymptotic (small h) expansion, and it can be argued that h_{n+1} offers a compromise between efficiency (choosing a large stepsize) and accuracy (satisfying the error criterion). If a step is rejected, then (1.8) can be used to determine a stepsize with which to repeat the step. Other techniques are also used in practice, but the precise details of stepsize changing after a rejection are not important for our analysis.

The main question that we address in this work is whether error control algorithms of the type described above will automatically suppress spuriosity. It must be emphasised that such error control is motivated by *local* quantities and $h_n \to 0$ expansions. In this work we are concerned with *long term* behaviour and *global* quantities. Here, the limit $t_n \to \infty$ is more relevant than the limit $h_n \to 0$. Although there seems to be a widely-held belief that 'error control suppresses spuriosity' (see, for example, [10]), to date this has not been rigorously established for general ERK methods and problems.

We mention that recent work by Stuart and Humphries [12] and Higham and Stuart [5] shows that local error control offers benefits for long-term computations with certain problems and methods. Our approach is less specific (and the results less precise), since we aim to gain insight into the behaviour of general ERK methods on general ODEs.

In the next section, we look at the existence of spurious fixed points in a variable stepsize setting. We show that such points are likely to arise whenever the individual spurious fixed point branches for the two formulas intersect. This corresponds to the intersection of two curves in $\mathbb{R}^{m+1}$. For scalar problems ($m = 1$) this scenario is not unlikely, and hence spuriosity cannot be ruled out. Section 3 examines the stability of spurious fixed points in the scalar case. We show that, in general, instability is inevitable for small tolerances, so spurious fixed points are unlikely to be seen in practice. We also provide numerical evidence that even when a spurious fixed point is stable, it is likely to have a basin of attraction that shrinks with τ. Sections 4 and 5 illustrate a technique for constructing "genuinely spurious" fixed points; that is, spurious fixed points that are stable and have a significant basin of attraction for small τ. Although such examples are extremely contrived, they illustrate the worst-case behaviour of standard error control schemes. Finally, in section 6, we summarise our conclusions.

2 Existence of spurious fixed points

2.1 The scalar case

Throughout this work, f in (1.1) is assumed to be C^1. In this subsection we assume that the ODE (1.1) is scalar ($m = 1$). The norm in (1.6) or (1.7) is taken to be the absolute value. This simplifies the analysis, without affecting our main conclusions. Ignoring step

rejections, the one-step recurrence given by (1.3) and (1.8) can then be written in the form

$$y_{n+1} = y_n + h_n \Phi(h_n, y_n), \tag{2.1}$$

$$h_{n+1} = \left(\frac{\widehat{\tau}_n}{h_n |\Phi(h_n, y_n) - \Psi(h_n, y_n)|} \right)^{1/q} h_n, \tag{2.2}$$

where $\widehat{\tau}_n = \theta^q \tau$ (independent of n) for EPS control, and $\widehat{\tau}_n = \theta^q \tau h_n$ for EPUS control.

A fixed point of this recurrence, that is, a solution where both h_n and y_n are constant, must satisfy

$$\Phi(h^*, y^*) = 0, \tag{2.3}$$

$$|\Psi(h^*, y^*)| = \widehat{\tau}/h^*. \tag{2.4}$$

The condition (2.3) forces (h^*, y^*) to be a (constant stepsize) fixed point of the main formula. The second condition, (2.4), which ensures that the stepsize remains constant on each step, forces (h^*, y^*) to be within $O(\tau)$ of a fixed point of the secondary formula. (Note that with such a solution the error criterion $\text{est}_{n+1} \leq \tau$ is satisfied, as required.)

Now consider the two ERK formulas separately. They may have branches of fixed points; that is, points (h, y) satisfying $\Phi(h, y) = 0$ and $\Psi(h, y) = 0$, respectively. These branches may intersect at some point $(\bar{h}, \bar{y})$. If so, then (by continuity) for small τ, moving away from $(\bar{h}, \bar{y})$ along the main branch $\Phi(h, y) = 0$ will generally perturb $|\Psi(h, y)|$ away from zero until (2.4) becomes satisfied. In general, we can move along the main branch in either of two directions. So, for a given small τ, we would expect there to exist two fixed points close to each intersection point.

How likely is it that two different ERK formulas will have spurious fixed point branches that intersect? In the case where $f(y)$ is a polynomial of degree d, it is mentioned in [1] (and can be easily seen from (1.2) and (1.3)) that for an s-stage ERK formula, the fixed points are precisely the real roots of a polynomial of degree d^s. Hence, a formula is likely to have many branches of spurious fixed points, and the potential for a spurious fixed point to be shared between two formulas seems high. The reference [1] plots bifurcation diagrams for several low order ERK formulas applied to quadratic and cubic polynomials f. Even in these cases, where s and d are small, by superimposing the figures it can be seen that fixed point branches intersect for many pairs of formulas. We have also conducted numerical experiments with several widely-used formulas and polynomial-like functions f. Our results support the tenet that simultaneous spurious fixed points are not rare.

As an example we consider a 4th and 5th order pair derived by Fehlberg, which is referred to as RKF45 by Lambert [8, page 185]. This pair has been used in many programs, including the influential `RKF` [11] and Matlab's `ode45.m` [9]. We apply the pair to the logistic problem $y'(t) = y(t)(1 - y(t))$. Figure 2.1 shows the results of a simple grid search for spurious fixed points of the individual formulas. The symbols '$\cdot$' and '$+$' are used for the 4th and 5th order formulas, respectively. The left-hand plot highlights the abundance of spurious fixed points for the two methods, with many fixed point branches lying in close proximity. A more detailed search on the domain $(h, y) \in [2.7, 2.8] \times [1.033, 1.042]$ reveals that the fixed point curves of the 4th and 5th order formulas intersect in this region and that the pair possesses a common spurious fixed point in the vicinity of $(h, y) = (2.76, 1.037)$. We point out that this stepsize lies

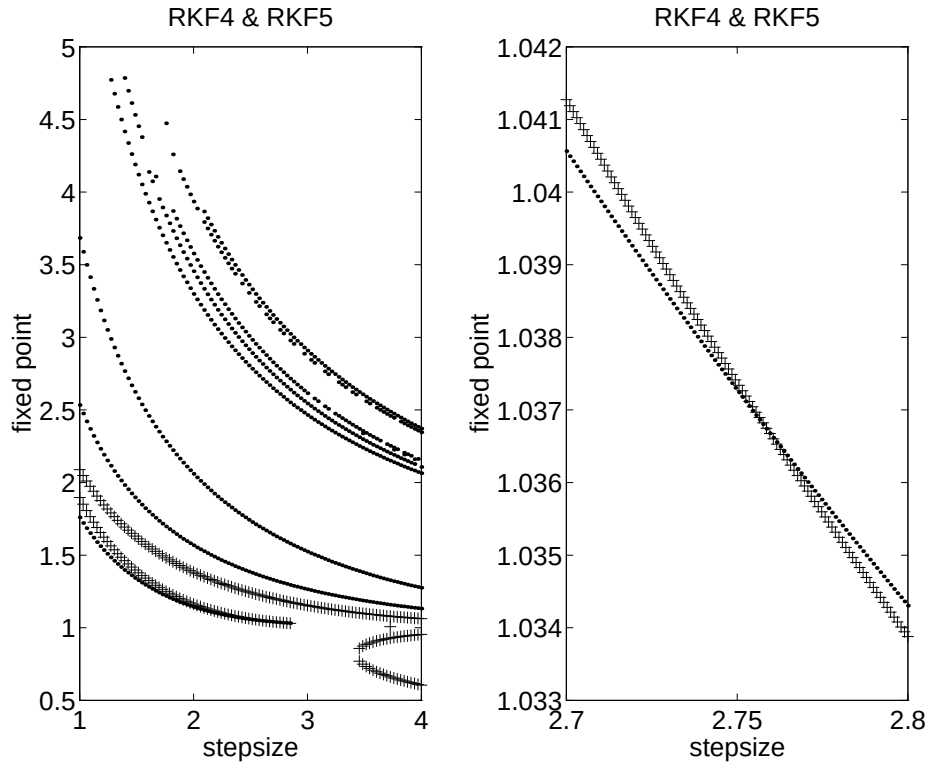

Figure 2.1: Spurious fixed points of 4th and 5th order Fehlberg method on logistic equation.

below the stability limit that arises (for either formula) from linearisation about the true, stable fixed point $y(t) \equiv 1$.

2.2 The system case

We consider now the case where $m > 1$ in (1.1). Our first observation is that a spurious fixed point for a scalar problem can be extended to a spurious fixed point for a system. For example, assuming for convenience that a p-norm is used to obtain est_{n+1}, if (h^*, y_1^*) is a spurious fixed point for the scalar problem $y'(t) = f_1(t, y(t))$, and if $f_2(y_2^*) = 0 \in \mathbb{R}$, then $(h^*, [y_1^*, y_2^*]^T)$ is a spurious fixed point for the system

$$\begin{aligned} y_1'(t) &= f_1(y_1(t)) + (y_2(t) - y_2^*)^2, \\ y_2'(t) &= f_2(y_2(t)). \end{aligned}$$

This idea can clearly be extended to higher dimensions, but it forces an extremely contrived type of coupling between components. We believe that for "genuine" systems of ODEs the existence of a spurious fixed point for small τ is highly unlikely.

When (1.1) is a system of ODEs the equations (2.3)–(2.4) for a fixed point of the adaptive algorithm become

$$\Phi(h^*, y^*) = 0, \tag{2.5}$$

$$\|\Psi(h^*, y^*)\| = \hat{\tau}/h^*, \tag{2.6}$$

with $\Phi, \Psi : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$. Now suppose that for all sufficiently small τ a spurious fixed point (h^*, y^*) exists, depending continuously on τ, with $h^* \to \bar{h} \neq 0$ and $y^* \to \bar{y}$ (finite)

as $\tau \to 0$. Then, by continuity, $(\bar{h}, \bar{y})$ must solve

$$\Phi(h, y) = 0, \tag{2.7}$$
$$\Psi(h, y) = 0. \tag{2.8}$$

The constraints (2.7)–(2.8) form $2m$ nonlinear equations in the $m + 1$ unknowns, and hence are overdetermined for $m > 1$. From a geometric point of view, generically, (2.7) and (2.8) each represent a curve in $\mathbb{R}^{m+1}$, which also suggests that, except for pathological cases, there is little chance of a solution to (2.7)–(2.8). Suppose, for example, that $(\bar{h}, \bar{y})$ solves (2.7). Then, if Φ_y is nonsingular at $(\bar{h}, \bar{y})$, the Implicit Function Theorem shows that there exist h_a and h_b such that for any $h_a < h < h_b$ there is a $y = y(h)$ for which $\Phi(h, y(h)) = 0$ (and, of course, $y(\bar{h}) = \bar{y}$). Now, by reducing the length of the interval (h_a, h_b) if necessary, we can ensure that the Jacobian of Φ_y is nonsingular at $(h, y(h))$ for each $h_a < h < h_b$. Hence, by the Inverse Function Theorem, if we regard h as fixed, the system

$$F(y) := \Phi(h, y) = 0$$

has a locally unique solution at $y(h)$. Overall we see that if $\Phi_y(\bar{h}, \bar{y})$ is nonsingular, there is a one-parameter family of solutions to (2.7) around $(\bar{h}, \bar{y})$. In particular, this condition holds when the stepsize is "stable" in the sense that $(\bar{h}, \bar{y})$ is a stable fixed point of the constant stepsize map (1.3). This follows because

$$\Re\{1 + \bar{h}\lambda\} < 0 \Rightarrow \Re\{\lambda\} < -1/\bar{h} < 0 \Rightarrow \lambda \neq 0,$$

for each eigenvalue λ of $\Phi_y(\bar{h}, \bar{y})$.

2.3 True fixed points

Although this work is concerned with *spuriousity*, we feel that it is worthwhile to mention briefly the behaviour of adaptive schemes around true fixed points. If y^* is a *true* fixed point of (1.1) then $f(y^*) = 0$ and it follows that $\Phi(h^*, y^*) = \Psi(h^*, y^*) = 0$ for any h^*. Hence, with $y_n = y^*$, we find $\text{est}_{n+1} = 0$ in (1.6) or (1.7). A zero error estimate is an "exception" that can be treated in various ways, but most programs would abandon the standard stepsize formula (1.8) and, over a sequence of steps, would increase the stepsize to some ceiling $h_{\max}$; so that $(h_{\max}, y^*)$ gave a fixed point. However, in practice, a code is unlikely to find y^* exactly, and it is more realistic to assume that y_n *approximates* the fixed point, and to consider the linearised problem. This scenario has been analysed, from a different viewpoint, by Hall [3]. In the scalar case, linearising about a stable, true fixed point y^*, produces the linear ODE

$$y'(t) = \lambda y(t), \quad \text{with} \quad \lambda := f_y(y^*) < 0, \tag{2.9}$$

which has a unique, stable fixed point $y(t) \equiv 0$. Hall showed that an adaptive ERK method will always admit a period one or two solution that is within $O(\tau)$ of this true fixed point. Further, a simple algebraic condition on the ERK coefficients determines the stability of the discrete solution. Recall that τ quantifies the level of accuracy required by the user, and it can thus be regarded as acceptable for the algorithm to have a fixed point that is within $O(\tau)$ of y^*. Analogous results for systems of ODEs can be found in [4].

3 Stability of spurious fixed points

The linear stability of the fixed point in (2.3)–(2.4) is determined by the spectral radius of the Jacobian of the map at h^*, y^*. The following Lemma exhibits the Jacobian.

Lemma 3.1 *The Jacobian of the map (2.1)–(2.2) at a fixed point h^*, y^* satisfying (2.3)–(2.4) has the form*

$$\begin{bmatrix} 1+h^*\Phi_y^* & h^*\Phi_h^* \\ \frac{-h^{*2}}{q\widehat{\tau}}\Gamma_y(h^*,y^*) & 1-\frac{1}{q}-\frac{h^{*2}}{q\widehat{\tau}}\Gamma_h(h^*,y^*) \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 1+h^*\Phi_y^* & h^*\Phi_h^* \\ \frac{-h^{*2}}{q\widehat{\tau}}\Gamma_y(h^*,y^*) & 1-\frac{h^{*2}}{q\widehat{\tau}}\Gamma_h(h^*,y^*) \end{bmatrix}, \tag{3.1}$$

for EPS and EPUS control, respectively. Here, $\Gamma = s^*(\Phi - \Psi)$, where $s^* = -\text{sign}\Psi(h^*, y^*)$.

Proof. First, consider the EPS case. Let $\omega = \theta^q\tau$. At (h^*, y^*) we have $\Phi(h^*, y^*) = 0$ and $|\Psi(h^*, y^*)| = \omega/h^*$. Letting $s^* = -\text{sign}\Psi(h^*, y^*)$, in the neighbourhood of h^*, y^* the map (2.1)–(2.2) is

$$\begin{aligned} y_{n+1} &= y_n + h_n\Phi(h_n, y_n), \\ h_{n+1} &= \left(\frac{\omega}{s^*h_n(\Phi(h_n, y_n) - \Psi(h_n, y_n))}\right)^{1/q} h_n. \end{aligned}$$

Letting $\Gamma = s^*(\Phi - \Psi)$, the map can be written

$$y_{n+1} = y_n + h_n\Phi(h_n, y_n), \tag{3.2}$$

$$h_{n+1} = \left(\frac{\omega}{\Gamma(h_n, y_n))}\right)^{1/q} h_n^{1-1/q}. \tag{3.3}$$

Note that $\Gamma(h^*, y^*) = |\Psi(h^*, y^*)| = \omega/h^*$.

We may write this map as $[y_{n+1}, h_{n+1}]^T = F([y_n, h_n]^T)$. The Jacobian at h^*, y^* can then be derived as follows.

For the $(1,1)$ and $(1,2)$ elements:

$$\frac{\partial F_1}{\partial y} = 1 + h^*\Phi_y^* \quad \text{and} \quad \frac{\partial F_1}{\partial h} = h^*\Phi_h^*.$$

The (2,1) element is

$$\begin{aligned} \frac{\partial F_2}{\partial y} &= \omega^{1/q}h_n^{1-1/q}\frac{-1}{q}\Gamma(h_n, y_n)^{-1-1/q}\Gamma_y(h_n, y_n) \\ &= \left(\frac{\omega}{h_n\Gamma(h_n, y_n)}\right)^{1/q} h_n\frac{-1}{q}\frac{1}{\Gamma(h_n, y_n)}\Gamma_y(h_n, y_n), \end{aligned}$$

which, at h^*, y^*, becomes

$$\frac{\partial F_2}{\partial y} = 1 \times h^*\frac{-1}{q}\frac{1}{\omega/h^*}\Gamma_y(h^*, y^*) \tag{3.4}$$

$$= \frac{-h^{*2}}{q\omega}\Gamma_y(h^*, y^*). \tag{3.5}$$

The (2,2) element is

$$\frac{\partial F_2}{\partial h} = \omega^{1/q}\left\{\frac{-1}{q}\Gamma(h_n,y_n)^{-1-1/q}\Gamma_h(h_n,y_n)h_n^{1-1/q} + \Gamma(h_n,y_n)^{-1/q}(1-\frac{1}{q})h_n^{-1/q}\right\} \quad (3.6)$$

$$= \left(\frac{\omega}{h_n\Gamma(h_n,y_n)}\right)^{1/q}\left\{\frac{-1}{q}\frac{\Gamma_h(h_n,y_n)}{\Gamma(h_n,y_n)}h_n + 1 - 1/q\right\}, \quad (3.7)$$

and hence at h^*, y^* we get

$$\frac{\partial F_2}{\partial h} = 1 \times \left\{\frac{-1}{q}\frac{h^*}{\omega}\Gamma_h(h^*,y^*)h^* + 1 - 1/q\right\}$$

$$= \frac{-h^{*2}\Gamma_h(h^*,y^*)}{q\omega} + 1 - 1/q.$$

For the EPUS case, at (h^*, y^*) we have $\Phi(h^*, y^*) = 0$ and $|\Psi(h^*, y^*)| = \omega$, and the map can be written

$$y_{n+1} = y_n + h_n\Phi(h_n, y_n), \quad (3.8)$$

$$h_{n+1} = \left(\frac{\omega}{\Gamma(h_n,y_n))}\right)^{1/q} h_n. \quad (3.9)$$

The $(1,1)$ and $(1,2)$ elements are the same as for the EPS case.
The (2,1) element is

$$\frac{\partial F_2}{\partial y} = \omega^{1/q}h_n\frac{-1}{q}\Gamma(h_n,y_n)^{-1-1/q}\Gamma_y(h_n,y_n)$$

$$= \left(\frac{\omega}{\Gamma(h_n,y_n)}\right)^{1/q} h_n\frac{-1}{q}\frac{\Gamma_y(h_n,y_n)}{\Gamma(h_n,y_n)},$$

so that at h^*, y^* we have

$$\frac{\partial F_2}{\partial y} = \frac{-h^*\Gamma_y(h^*,y^*)}{q\omega}. \quad (3.10)$$

The (2,2) element is

$$\frac{\partial F_2}{\partial h} = \omega^{1/q}\left\{\frac{-1}{q}\Gamma(h_n,y_n)^{-1-1/q}\Gamma_h(h_n,y_n)h_n + \Gamma(h_n,y_n)^{-1/q}\right\}$$

$$= \left(\frac{\omega}{\Gamma(h_n,y_n)}\right)^{1/q}\left\{\frac{-1}{q}\frac{\Gamma_h(h_n,y_n)}{\Gamma(h_n,y_n)}h_n + 1\right\},$$

which, at h^*, y^*, gives

$$\frac{\partial F_2}{\partial h} = 1 - \frac{h^*\Gamma_h(h^*,y^*)}{q\omega}. \quad (3.11)$$

Using $\hat{\tau} = \omega h_n$ in (3.10) and (3.11) completes the proof. ∎

We are interested in spurious fixed points that persist for arbitrarily small τ; that is, where the fixed point converges to a limiting case $\bar{y}, \bar{h}$ at $\tau = 0$. The work of Humphries [6] makes it reasonable to assume that $\bar{h} \neq 0$. (Since, for smooth f, a spurious fixed point that persists as $h \to 0$ must become unbounded.) The following theorem deals with this scenario, showing that in, general, a fixed point cannot remain stable as $\tau \to 0$.

Theorem 3.1 *Suppose that a fixed point h^*, y^* satisfying (2.3)–(2.4) exists for all sufficiently small τ, and suppose that h^*, y^* depend continuously on τ with*

$$h^* \to \bar{h} \neq 0, \quad y^* \to \bar{y} \quad \text{finite}, \quad \text{as} \quad \tau \to 0. \tag{3.12}$$

Then a necessary condition for h^*, y^* to be linearly stable for small τ is

$$\Psi_h(\bar{h}, \bar{y}) - \Phi_h(\bar{h}, \bar{y}) = \Psi_y(\bar{h}, \bar{y}) - \Phi_y(\bar{h}, \bar{y}) = 0. \tag{3.13}$$

Proof. Consider the EPS case. Linear stability requires the spectral radius of the 2×2 Jacobian, J_{EPS}, to be strictly less than one. Using the Routh-Hurwitz condition [8, page 14] this may be written

$$\det(J_{\text{EPS}}) < 1 \quad \text{and} \tag{3.14}$$

$$|\text{trace}(J_{\text{EPS}})| < 1 + \det(J_{\text{EPS}}). \tag{3.15}$$

Using (3.14) in (3.15), necessary conditions for stability are

$$\det(J_{\text{EPS}}) < 1 \quad \text{and} \tag{3.16}$$

$$|\text{trace}(J_{\text{EPS}})| < 2. \tag{3.17}$$

From Lemma 3.1 these conditions are

$$\left(1 + h^*\Phi_y^*\right)\left(1 - 1/q - \frac{h^{*2}s^*\{\Psi_h(h^*, y^*) - \Phi_h(h^*, y^*)\}}{q\widehat{\tau}}\right)$$

$$+\frac{h^{*2}}{q\widehat{\tau}}s^*\{\Psi_y(h^*, y^*) - \Phi_y(h^*, y^*)\}h^*\Phi_h^* < 1 \quad \text{and} \tag{3.18}$$

$$\left|2 - 1/q + h^*\Phi_y^* - \frac{h^{*2}s^*\{\Psi_h(h^*, y^*) - \Phi_h(h^*, y^*)\}}{q\widehat{\tau}}\right| < 2. \tag{3.19}$$

Consider (3.19), under our assumptions (3.12), as $\tau \to 0$. This condition cannot hold unless

$$\Psi_h(h^*, y^*) - \Phi_h(h^*, y^*) = O(\tau). \tag{3.20}$$

(Otherwise, precisely one term in (3.19) is unbounded as $\tau \to 0$.) Using this in (3.18) it follows that we must have

$$\Psi_y(h^*, y^*) - \Phi_y(h^*, y^*) = O(\tau). \tag{3.21}$$

Combining (3.20) and (3.21), a necessary condition for stability as $\tau \to 0$ is

$$\Psi_h(\bar{h}, \bar{y}) - \Phi_h(\bar{h}, \bar{y}) = 0 = \Psi_y(\bar{h}, \bar{y}) - \Phi_y(\bar{h}, \bar{y}),$$

giving the result.

A similar proof can be used for the EPUS case.

∎

Note that, in particular, the condition (3.13) forces

$$\frac{\Phi_y(\bar{h},\bar{y})}{\Phi_h(\bar{h},\bar{y})} = \frac{\Psi_y(\bar{h},\bar{y})}{\Psi_h(\bar{h},\bar{y})}. \tag{3.22}$$

The relation (3.22) has a geometrical interpetation—it implies that the the two branches of fixed points for the corresponding fixed-stepsize methods must intersect tangentially. Hence, although any intersection can give rise to a fixed point, only the pathological case of a *tangential* intersection can produce a fixed point that is stable for small τ.

Note also that (3.13) implies that if we expand $\Psi - \Phi$ about $(\bar{h},\bar{y})$ then the first nonzero terms have order 2 in $h^* - \bar{h}$ and $y^* - \bar{y}$. Hence the difference $\Psi - \Phi$ is flat in the sense that the gradient is zero at $(\bar{h},\bar{y})$. In later sections, where we construct stable spurious fixed points, we will force the ultimate flatness—by using piecewise constant functions, all terms in the expansion will be zero.

In the previous section, we saw that the formulas in the RKF45 pair share a common spurious fixed point on the logistic equation, and hence an error controlled algorithm will admit spurious fixed points for small τ. For illustration, we consider the non-extrapolation mode, that is, with the 4th order formula advancing the solution, using error-per-step control and a safety factor of $\theta = .8$. The left-hand plot of Figure 3.1 gives the stepsize at

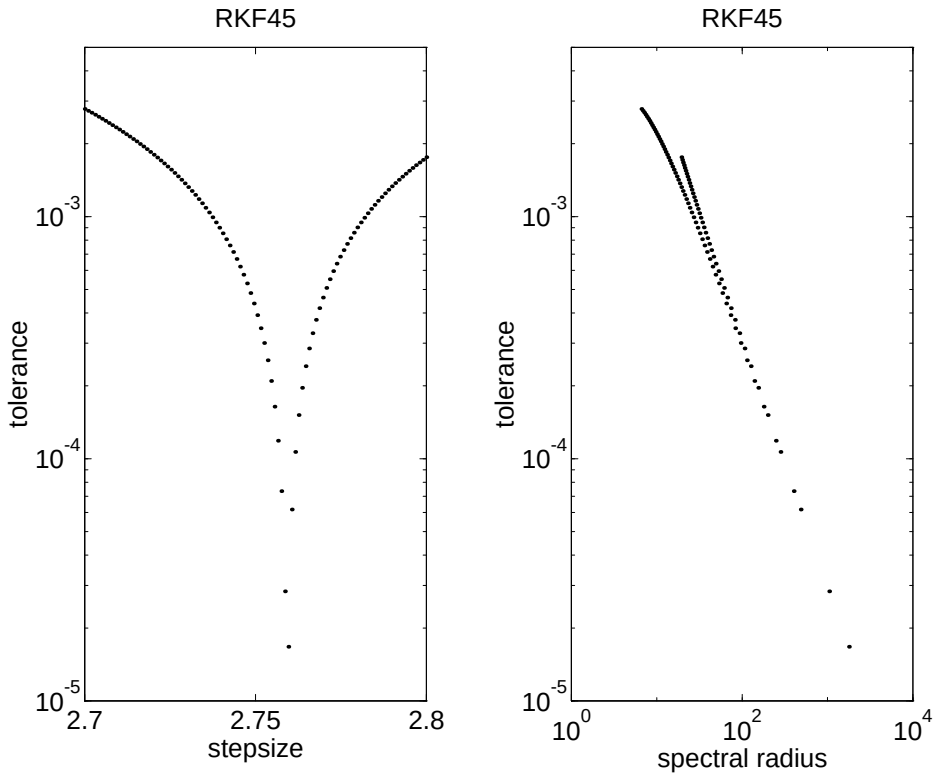

Figure 3.1: Fixed points and spectral radii for the RKF45 method applied to the logistic equation.

a spurious fixed point against τ. We see that for each small τ, two spurious fixed points exist, one on either side of the point $\bar{h}$ where the intersection in Figure 2.1 appears. Further, as $\tau \to 0$ the stepsizes tend to $\bar{h}$. This agrees with the theory discussed in the previous section. In the right-hand plot, the spectral radius of the corresponding Jacobian is given. We see that in all cases the fixed point is unstable, and the spectral

radius increases like $1/\tau$ as $\tau \to 0$. This is consistent with the presence of the $1/\tau$ terms in the expression for the Jacobian in Lemma 3.1.

It is important to emphasise that Theorem 3.1 applies when $\tau \to 0$; it is possible for a method to have a stable spurious fixed for some fixed value of τ. We illustrate this with the second and third order pair used in Matlab's `ode23.m` routine. (The coefficients of this pair are given at the start of section 5, along with a proof that a fixed point cannot remain spurious as $\tau \to 0$.) Figure 3.2 shows the spurious fixed points of the two formulas in constant-stepsize mode on the logistic equation. The right-hand picture zooms in on a particular region for the third order formula. Taking each formula in turn, a fixed

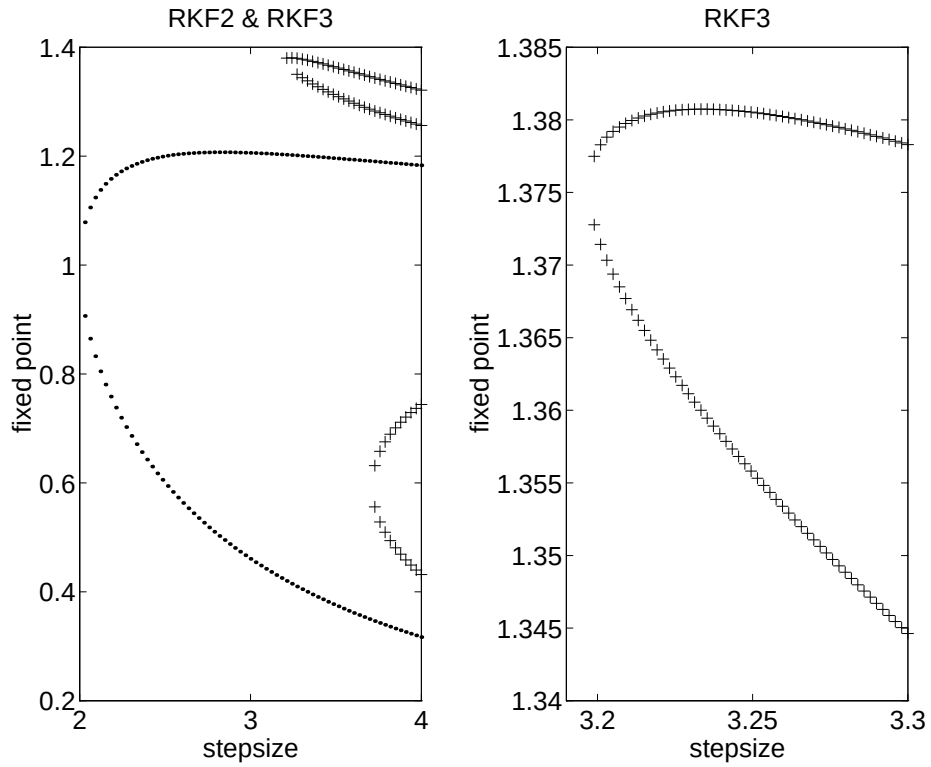

Figure 3.2: Spurious fixed points of 2nd (.) and 3rd (+) order formula on logistic equation with constant stepsize.

point of the corresponding error-controlled algorithm can be constructed by choosing τ so that (2.4) holds; in other words, we fix the tolerance so that the stepsize formula reproduces the required stepsize. Figures 3.3 and 3.4 illustrate this approach in the case of non-extrapolation and extrapolation, respectively. The left-hand pictures show the resulting tolerances and the right-hand pictures give the spectral radius of the Jacobian of the map. The '+' symbol marks fixed points that are stable. For the second order formula in Figure 3.3 stable fixed points exist, but only in a region where y^* approaches the true fixed point. With the third order formula, however, it can be seen that genuinely spurious, stable, fixed points exist, along the branch highlighted in the right-hand picture of Figure 3.2. The corresponding value of τ is approximately 2.

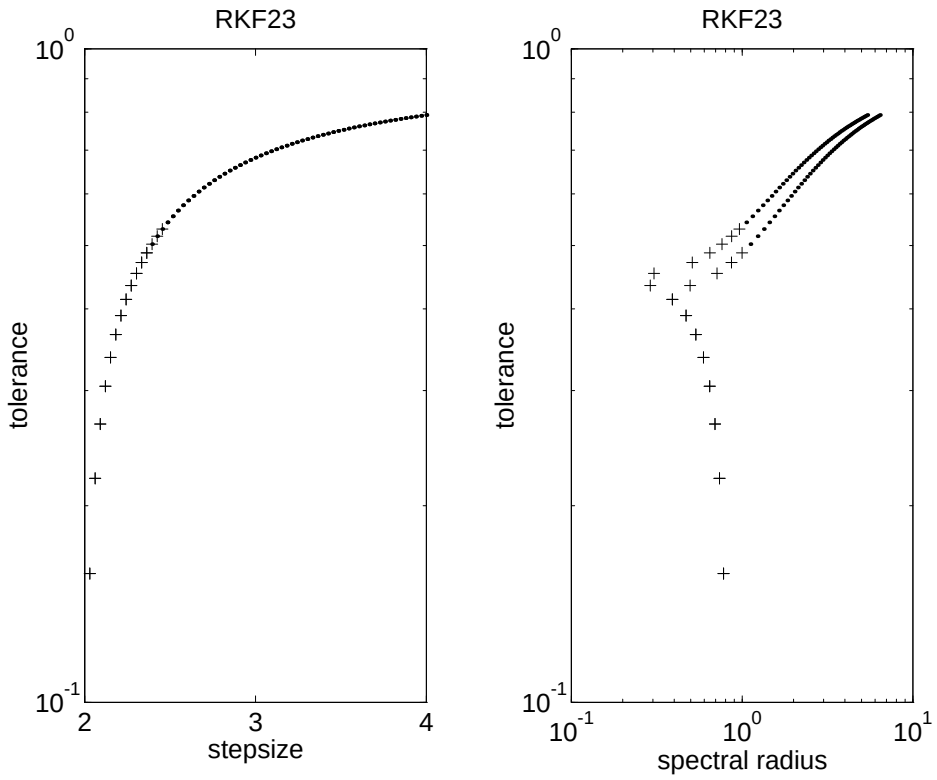

Figure 3.3: Fixed points and spectral radii for the 2(3) method applied to the logistic equation.

4 Examples of Stable Spuriosity

Our aim in this section is to construct examples where stable, spurious fixed points exist for small τ. This motivates the general analysis in section 5. In both sections we restrict attention to scalar problems ($m = 1$) but, as discussed in section 2.2, examples where $m > 1$ can be built up from the $m = 1$ case.

4.1 A Polynomial Example

The ERK pair we choose comprises the 2nd order Improved Euler Method and another 2-stage method which is 1st order. Here, the increment functions (1.3) and (1.5) are

$$\begin{aligned} \Phi(h,y) &= \frac{1}{2}(f(y) + f(y + hf(y))), \\ \Psi(h,y) &= \frac{1}{3}(f(y) + 2f(y + \frac{3}{5}hf(y))). \end{aligned}$$

We remark that in contrast to usual error control algorithms, which normally use embedded formulas, our example has distinct stage values for the primary and secondary methods. This is purely a matter of convenience—we show later that the use of embedded ERK formulas does not prevent stable spuriosity. We assume that EPS control (1.6) is used, with $\theta = .9$ in the stepsize selection formula (1.8).

We introduce the notation $x = y + hf(y)$, $z = y + 3hf(y)/5$ to denote the primary and secondary stage values. Then, from (2.3)–(2.4), a fixed point of the Runge-Kutta

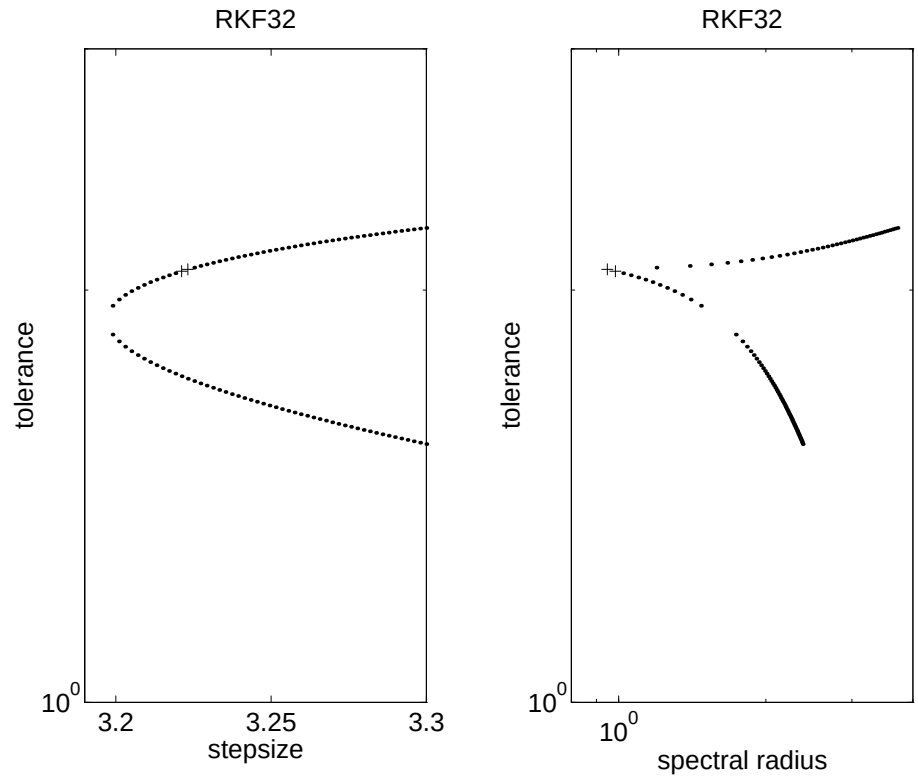

Figure 3.4: Fixed points and spectral radii for the 3(2) method applied to the logistic equation.

pair satisfies

$$\Phi(h^\star, y^\star) = \frac{1}{2}(f(y^\star) + f(x^\star)) = 0, \tag{4.1}$$

$$h^\star|\Psi(h^\star, y^\star)| = \frac{1}{3}h^\star|f(y^\star) + 2f(z^\star)| = \theta^q\tau = \widehat{\tau}. \tag{4.2}$$

Now, let $\epsilon = \widehat{\tau}/h^\star$ and suppose $\Psi(h^\star, y^\star) < 0$. If we make the normalization $f(y^\star) = -1$, which ensures that $y^\star$ is genuinely spurious, then the conditions to be satisfied are

$$x^\star = y^\star - h^\star, \quad z^\star = y^\star - \tfrac{3}{5}h^\star, \quad f(x^\star) = 1 \quad \text{and} \quad f(z^\star) = \tfrac{1}{2}(1 - 3\epsilon).$$

Any function in (1.1) satisfying these criteria will necessarily possess a spurious fixed point at $(h^\star, y^\star)$. We note that the location of $y^\star$ has not yet been fixed.

Thus far we have ignored the stability of the spurious fixed point. We illustrated in section 3 how the stability depends upon the partial derivatives of the increment functions. If we ensure that $\partial\Phi(h, y)/\partial h = 0$ and $\partial\Psi(h, y)/\partial h = 0$ at the fixed point then the terms in the Jacobian depending upon the reciprocal of the error tolerance do not contribute to the spectral radius and the fixed point may be stable for small τ. Therefore, we ask for f to have its turning points coincident with the stage values of the method. Construction of such a function is trivial.

Locating the spurious fixed point at $(h^\star, y^\star) = (2, 5/2)$ and taking the function parameter $\epsilon = 4/10000$ yields the interpolating polynomial

$$f(y) = -\frac{186079}{98304} + \frac{777335}{49152}y - \frac{376763}{12288}y^2 + \frac{158407}{6144}y^3 - \frac{60475}{6144}y^4 + \frac{4235}{3072}y^5, \tag{4.3}$$

where the stage and function values are

$$x^\star = 1/2, \quad z^\star = 13/10, \quad f(x^\star) = 1 \quad \text{and} \quad f(z^\star) = 2497/5000 = \tfrac{1}{2}(1 - 3\tfrac{4}{10000});$$

see Figure 4.1. Note the presence of a true, stable, fixed point close to $y = 2$.

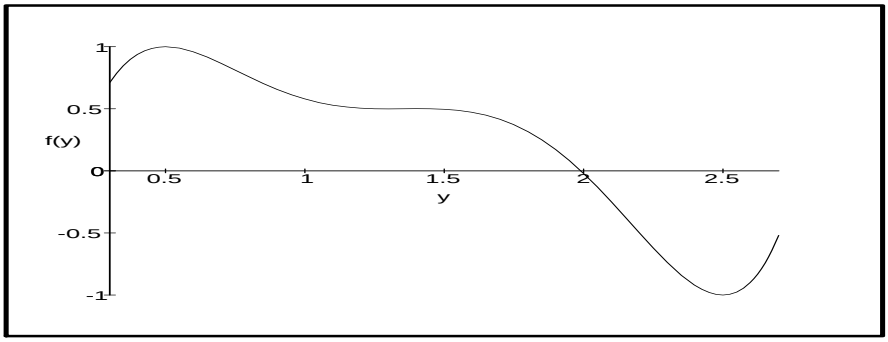

Figure 4.1: Graph of Polynomial (4.3).

The spurious fixed point constructed by this process is linearly stable. However, its basin of attraction is very small—perturbations of $O(\tau)$ lead to attraction to the true fixed point. Figure 4.2 shows the results of taking 100 steps on the domain of initial data

$$(h_0, y_0) \in [1.999, 2.001] \times [2.499, 2.501]$$

with the error tolerance $\tau = 0.0009877$. The lower pictures are contour plots, with the contour heights chosen at regular intervals between the extreme values. It is clear that $(h^\star, y^\star)$ has a tiny basin of attraction.

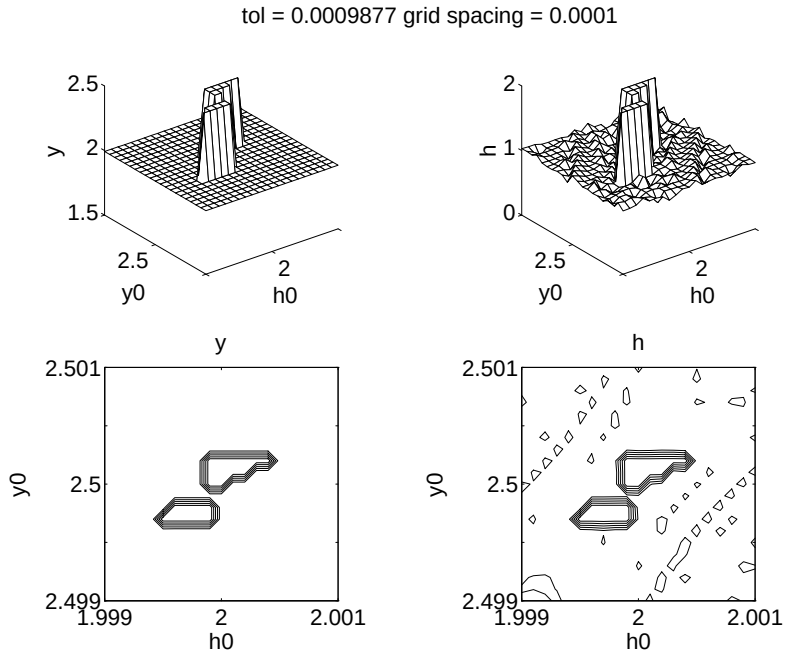

Figure 4.2: Basin of Attraction for the Spurious Equilibrium Point of Polynomial (4.3).

We have conducted other numerical experiments with similar results. Our experience suggests that with polynomial-like functions f, even when a spurious fixed point is stable

its basin of attraction is small and typically shrinks with τ. This is caused by the presence of the $1/\tau$ terms in the Jacobian (see Lemma 3.1). It is possible to force the appropriate elements in the Jacobian to be zero at $(h^\star, y^\star)$, but away from this point the elements typically increase rapidly. Hence, the Jacobian is likely to be "repelling" except in a very small neighbourhood of $(h^\star, y^\star)$.

The significant feature determining the size of the basin of attraction is the function curvature at the spurious fixed point and stage values. If the curvature is large then small perturbations of the data may lead to the occurence of large terms in the Jacobian and instability. To ensure a significant basin of attraction, we turn our attention to a locally piecewise constant function.

4.2 Genuine Spuriosity

Consider the function

$$f(y) = \begin{cases} 1 & : y \in (-\infty, y_1] \\ (y - 1 - \mu)m & : y \in [y_1, y_2] \\ -m\mu & : y \in [y_2, y_3] \\ (y-2)m & : y \in [y_3, y_4] \\ -1 & : y \in [y_4, \infty) \end{cases} \tag{4.4}$$

where the parameters μ and $\epsilon = \widehat{\tau}/h^\star$ satisfy

$$0 < \mu \leq 1 \quad 0 < \epsilon < 1/3,$$

the gradient of the sloping line segments is defined as

$$m = -\frac{1}{2\mu}(1 - 3\epsilon),$$

and the intervals are chosen so that

$$y_1 = 1 + \mu + 1/m, \quad y_2 = 1, \quad y_3 = 2 - \mu, \quad y_4 = 2 - 1/m.$$

With our choice of parameters (1.1) possesses a unique stable true fixed point at $y = 2$, see Figure 4.3. Given any initial condition, $y(t)$ will converge to this true fixed point as $t \to \infty$.

Conditions (2.3)–(2.4) show that a fixed point of the Runge-Kutta pair must be a fixed point of the primary method and must, to within a small perturbation, be a fixed point of the secondary method. We proceed by considering the fixed points of the primary and perturbed secondary methods in turn.

At a fixed point of the primary method the stage value, $x^\star$, satisfies

$$f(x^\star) = -f(y^\star) \tag{4.5}$$

for a certain stepsize, h. In the interval $y^\star \in (-\infty, y_1]$ the solutions for the primary formula are given by

$$f(y^\star) = 1, \quad f(x^\star) = -1,$$

where $x^\star = y^\star + hf(y^\star) = y^\star + h$. Considering each interval

$$x^\star \in (-\infty, y_1], \quad x^\star \in [y_1, y_2], \quad x^\star \in [y_2, y_3], \quad x^\star \in [y_3, y_4] \quad \text{and} \quad x^\star \in [y_4, \infty),$$

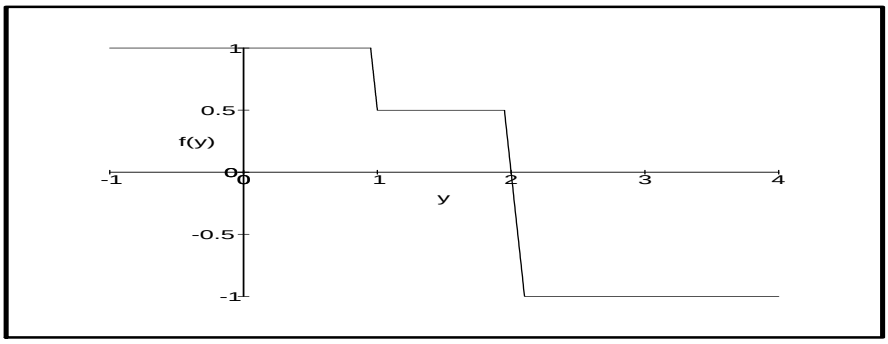

Figure 4.3: The Example Function with $\mu = 0.05$ and $\epsilon = 0.0004$.

in turn, we find that the stage value satisfies $f(x^\star) = -1$ when $x^\star \in [y_4, \infty)$. Rearranging and substituting for the stage value gives the fixed stepsize as

$$h \in [y_4 - y^\star, \infty).$$

Similarly for the interval $y^\star \in [y_1, y_2]$ the solutions of the primary method are given by

$$f(y^\star) = (y^\star - 1 - \mu)m, \quad f(x^\star) = -(y^\star - 1 - \mu)m,$$

where $x^\star = y^\star + (y^\star - 1 - \mu)mh$. Again considering each subinterval we obtain the fixed stepsize as

$$h = \frac{(3 + \mu - 2y^\star)}{(y^\star - 1 - \mu)m}.$$

Continuing in this manner for the remaining intervals we find the stepsizes associated with the fixed solutions are

$$\left.\begin{array}{lll} y^\star \in [y_2, y_3], & x^\star \in [y_3, y_4], & h = \frac{(y^\star - 2 - \mu)}{m\mu}. \\ y^\star = 2, & x^\star = 2, & \forall h, \\ y^\star \in [y_3, 2 + \mu], & x^\star \in [y_3, 2 + \mu], & h = -2/m, \\ y^\star \in [2 + \mu, y_4], & x^\star \in [y_1, y_2], & h = \frac{(3 + \mu - 2y^\star)}{(y^\star - 2)m}. \\ y^\star \in [y_4, \infty), & x^\star \in (-\infty, y_1], & h \in [y^\star - 1, \infty). \end{array}\right\}$$

We now have the complete fixed point diagram for the primary method for all stepsizes. Once the error tolerance is specified, the fixed stepsize for the solution is defined as $h^\star = \hat{\tau}/\epsilon$. For the choice $\mu = 0.05$, $\epsilon = 0.0004$, Figure 4.4 shows the large regions where the primary increment function is zero.

Now we consider the secondary method, which must satisfy (2.4). Since $\epsilon = -\Psi(h^\star, y^\star)$, the stage value $z^\star$ must satisfy

$$f(z^\star) = -\frac{1}{2}(f(y^\star) + 3\epsilon) = -\frac{1}{2}(f(y^\star) + 2\mu m + 1)$$

for a certain stepsize, h. As for the primary method we consider separate subintervals.

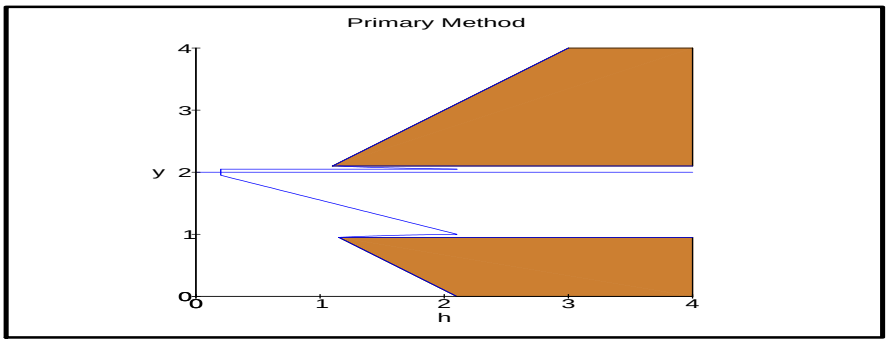

Figure 4.4: Fixed points for the Primary Method.

This gives the solution, stage value and stepsize of the perturbed secondary method as

$$\left.\begin{array}{lll} y^\star \in (-\infty, y_1], & z^\star \in [y_3, y_4], & h = \frac{5}{3}(2 - y^\star - \mu - 1/m), \\ y^\star \in [y_1, y_2], & z^\star \in [y_1, y_2], & h = \frac{5}{6(y^\star - 1 - \mu)m}(-3y^\star + 3 + \mu - 1/m), \\ y^\star \in [y_1, y_2], & z^\star \in [y_3, y_4], & h = \frac{5}{6(y^\star - 1 - \mu)m}(-3y^\star + 5 - \mu - 1/m), \\ y^\star \in [y_2, y_3], & z^\star \in [y_3, y_4], & h = \frac{5}{6m\mu}(2y^\star - 4 + \mu + 1/m), \\ y^\star \in [y_3, 2) \cup (2, y_4], & z^\star \in [y_3, 2) \cup (2, y_4], & h = \frac{5}{6(y^\star - 2)m}(-3y^\star + 6 - 2\mu - 1/m), \\ y^\star \in [y_4, \infty), & z^\star \in [y_2, y_3], & h \in [\frac{5}{3}(y^\star - y_3), \frac{5}{3}(y^\star - y_2)]. \end{array}\right\}$$

We now have the complete "fixed point" diagram for the perturbed secondary method, see Figure 4.5.

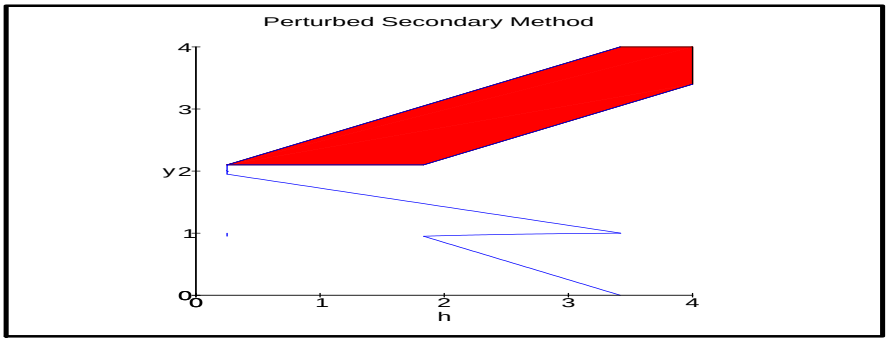

Figure 4.5: Fixed points for the Perturbed Secondary Method.

It is worthy of note that the fixed point diagrams in Figure 4.4 and Figure 4.5 contain two-dimensional *regions* in addition to the one-dimensional *branches* that are usually seen. This is because f is locally piecewise constant. Any point $(h^\star, y^\star)$ that appears in both figures is a fixed point of the corresponding error-controlled method with $\tau = h^\star \epsilon$.

Since the two figures generally intersect in regions (rather than at isolated points) it is clear that for a particular $h^\star$, there is generally an interval $[y_{\min}, y_{\max}]$ for which $(h^\star, y^\star)$ is a fixed point. Hence, perturbing $y^\star$ so that it remains in this interval still gives a fixed point with stepsize $h^\star$. Because of the locally piecewise constant nature of f, it can then be shown that given $y^\star \in [y_{\min}, y_{\max}]$ there is an interval $[h_{\min}, h_{\max}]$ such that

$$(h_0, y_0) \in [h_{\min}, h_{\max}] \times [y_{\min}, y_{\max}] \Rightarrow h_n \to h^\star (\text{monotonically}), y_n \equiv y_0,$$

as $n \to \infty$. (We do not present the details here—a general proof is given in the next section.) Hence there is a non-trivial, connected region of initial conditions that produce spurious solutions.

Taking the same function parameters used to produce the fixed point diagrams above; that is, $\mu = 0.05$, $\epsilon = 0.0004$, and choosing the error tolerance to be $\tau = 0.001 = \widehat{\tau}/0.9^2$ gives the fixed stepsize, $h^\star = \widehat{\tau}/|\epsilon| = 2.025$. Figure 4.6 graphs the results of applying the error controlled Runge-Kutta method 100 times on the domain of initial data $(h_0, y_0) \in [1,3] \times [2,4]$.

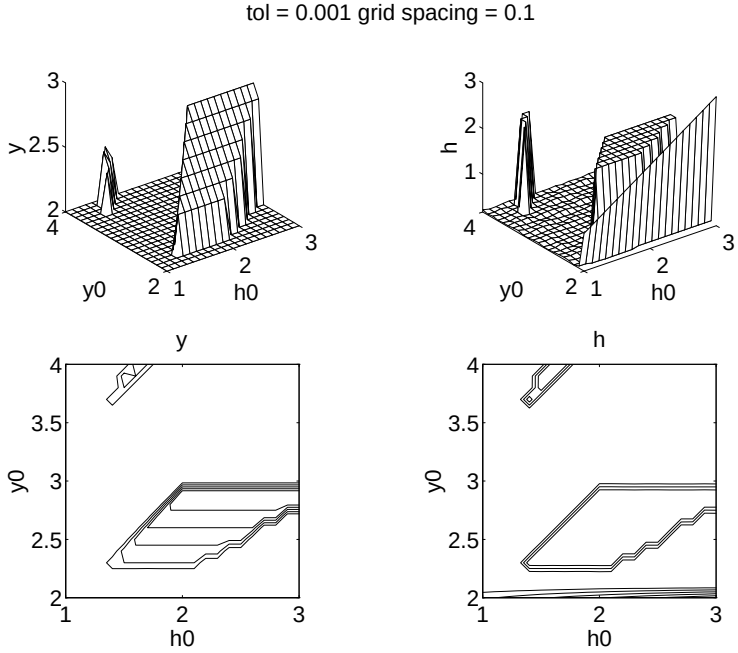

Figure 4.6: Basin of Attraction for the Spurious Equilibrium Point of Equation (4.4).

There are several features to note in Figure 4.6. First, there are two large regions of spurious fixed points shown in the left-hand plots. The larger of the two regions corresponds with where the fixed point diagrams of the two methods overlap. As the theory predicts, the spurious solution here is equal to the initial value, and the stepsize chosen by the error-control mechanism has converged monotonically to $h^\star = 2.025$.

We also note a second region of spurious solutions for initial data near $(h_0, y_0) = (3/2, 4)$. Again the stepsize calculated by the error-control algorithm converges to $h^\star = 2.025$. These solutions are produced by another feature of the stepsize selection algorithm—stepsize rejection. Here a stepsize is rejected because the error estimate is greater than the specified error tolerance and a smaller stepsize is then used. The new stepsize takes the iterates inside the main part of the basin of attraction.

The final feature to notice appears in the right-hand plots of Figure 4.6 where the stepsize for initial data $(h_0, y_0) = (h, 2)$ is equal to its initial value. Here, since $y \equiv 2$ is a true fixed point, the error estimate is identically zero, and the stepsize is arbitrarily left unchanged by the stepsize selection algorithm.

5 A Technique for Producing Genuine Spuriosity

Our aim now is to develop a general technique for constructing stable spurious fixed points that exist for small τ. (As in the previous section, we consider scalar problems only.) We begin with an example showing that this goal cannot always be achieved. The ERK formula used in the built-in `ode23.m` function of Matlab [9] has coefficients $a_{21} = 1$, $a_{31} = 1/4$, $a_{32} = 1/4$, $b = [1/6, 1/6, 4/6]$, $\widehat{b} = [1/2, 1/2, 0]$. Suppose that a fixed point exists for some function f. We allow f to vary with τ, and we suppose that f is a C^1 function of y and τ. We will denote the fixed point by h^*, X_1^* and let $X_i^* = X_1^* + h\sum_{j=1}^{i-1} a_{ij} f_j^*$ and $f_i^* = f(X_i^*)$. From (2.3)–(2.4) it follows that

$$\begin{aligned} f_1^* + f_2^* + 4f_3^* &= 0, \\ f_1^* + f_2^* &= O(\tau), \end{aligned}$$

giving $X_3^* = X_1^* + h^*(f_1^* + f_2^*)/4 = X_1^* + O(\tau)$. Therefore, by continuity, $f_1^* = f_3^* + O(\tau)$. Since $f_3^* = (f_1^* + f_2^*)/4 = O(\tau)$, we find that $f(X_1^*) = O(\tau)$. This shows that the fixed point cannot be truly spurious: $f(X_1^*)$ is zero to within order τ.

In general, the equations that a fixed point must satisfy are

$$\begin{aligned} X_2^* &= X_1^* + h^* a_{21} f_1^*, \\ X_3^* &= X_1^* + h^* a_{31} f_1^* + h^* a_{32} f_2^*, \\ &\vdots \\ X_s^* &= X_1^* + h^* \sum_{j=1}^{s-1} a_{sj} f_j^*, \end{aligned} \tag{5.1}$$

and

$$b_1 f_1^* + b_2 f_2^* + \ldots + b_s f_s^* = 0, \tag{5.2}$$

along with

$$\widehat{b}_1 f_1^* + \widehat{b}_2 f_2^* + \ldots + \widehat{b}_s f_s^* = \pm\widehat{\tau}/h^*, \tag{5.3}$$

where $\widehat{\tau} = \theta^q \tau$ for EPS and $\widehat{\tau} = \theta^q \tau h^*$ for EPUS. Generally, any function f for which $f(X_i^*) = f_i^*$ will then give rise to this fixed point. We will concentrate on the case where f is *locally piecewise constant*; that is, $f(x) = f_i^*$ for $x \in I_i$, where I_i is a connected, closed subinterval containing X_i^*. In order for such a function to exist, it is clearly necessary that $X_i^* \neq X_j^*$ whenever $f_i^* \neq f_j^*$. A more restrictive, but simpler, condition is that the X_i^* are distinct. We also require $f_1^* \neq 0$ for the fixed point to be spurious. This motivates the following definition.

Definition: A solution $h^*, \{X_i^*, f_i^*\}_{i=1}^s$ of (5.1), (5.2) and (5.3) is said to be **S-acceptable** if $\{X_i^*\}_{i=1}^s$ are distinct and $f_1^* \neq 0$.

The next lemma shows that an S-acceptable solution gives rise to a stable spurious fixed point, and provides a lower bound on the size of the basin of attraction.

Lemma 5.1 *Given an S-acceptable solution to (5.1), (5.2) and (5.3), suppose f is a locally piecewise constant function satisfying $f(x) = f_i^*$ for $x \in I_i$, where the subintervals I_i are disjoint with $X_i^* \in I_i$. Let*

$$f_{\max} := \max_i\{|f_i^*|\}, \quad \text{and} \quad \delta := \min_i\{\inf|\beta| : X_i^* + \beta \notin I_i\}. \tag{5.4}$$

(Note, δ can loosely be regarded as the "width" of the subintervals.)

Suppose that on a particular step we have a numerical solution $\widehat{X}_1 = X_1^* + \epsilon$ and a stepsize $\widehat{h} = h^* + \gamma > 0$.

With EPS control, if

$$|\epsilon| + |\gamma| s f_{\max} \max\{|a_{ij}|\} < \delta, \tag{5.5}$$

$$\gamma < \frac{h^*(1-\theta^q)}{\theta^q}, \tag{5.6}$$

then the pair $(\widehat{X}_1, \widehat{h})$ lies in the basin of attraction of a spurious fixed point. More precisely, $y_n = \widehat{X}_1$ for all n, and h_n converges monotonically to h^*.

With EPUS control, if (5.5) holds then the pair $(\widehat{X}_1, \widehat{h})$ lies in the basin of attraction of a spurious fixed point. More precisely, y_n and h_n remain constant for all n.

Proof. The first part of the proof shows that with the constraint (5.5) all the $f(X_i)$ values remain the same, so that the numerical solution is unchanged at the end of the step.

We start the step with $y_n = \widehat{X}_1$ and $h_n = \widehat{h}$. The second stage value is

$$\widehat{X}_2 = \widehat{X}_1 + \widehat{h} a_{21} f(\widehat{X}_1).$$

Since $|\epsilon| < \delta$ we have $f(\widehat{X}_1) = f(X_1^*)$, so

$$\widehat{X}_2 = X_1^* + \epsilon + (h^* + \gamma) a_{21} f(X_1^*).$$

Hence

$$\widehat{X}_2 - X_2^* = \epsilon + \gamma a_{21} f(X_1^*),$$

and so

$$|\widehat{X}_2 - X_2^*| \le |\epsilon| + |\gamma||a_{21}||f(X_1^*)| < \delta.$$

So we have $f(\widehat{X}_2) = f(X_2^*)$.

Proceeding by induction, suppose $f(\widehat{X}_j) = f(X_j^*)$, for $j = 1, 2 \ldots, i-1$. Then

$$\begin{aligned} \widehat{X}_i &= X_1^* + \epsilon + (h^* + \gamma) \sum_{j=1}^{i-1} a_{ij} f(\widehat{X}_j) \\ &= X_1^* + \epsilon + (h^* + \gamma) \sum_{j=1}^{i-1} a_{ij} f(X_j^*). \end{aligned}$$

So,

$$\widehat{X}_i - X_i^* = \epsilon + \gamma \sum_{j=1}^{i-1} a_{ij} f(X_j^*),$$

giving

$$|\widehat{X}_i - X_i^*| \le |\epsilon| + |\gamma|(i-1) f_{\max} \max\{|a_{ij}|\} < \delta.$$

So $f(\widehat{X}_i) = f(X_i^*)$. Since $\sum_{i=1}^{s} b_i f(X_i^*) = 0$, it follows that $y_{n+1} = \widehat{X}_1$.

The second part of the proof deals with the error control. We must show that the step will be accepted and the stepsize generated for the next step remains in a suitable neighbourhood of h^*.

Consider EPS control. The error estimate at the fixed point satisfies

$$\mathrm{est}_{n+1} := |\, h^* \sum_{i=1}^{s} \hat{b}_i f(X_i^*) \,| = \hat{\tau} = \theta^q \tau.$$

From the first part of the proof, we know that the $f(X_i)$ values are unchanged by the perturbation, hence the perturbed error estimate satisfies

$$\widehat{\mathrm{est}}_{n+1} := |\, (h^* + \gamma) \sum_{i=1}^{s} \hat{b}_i f(X_i^*) \,| = (1 + \frac{\gamma}{h^*})\hat{\tau}. \tag{5.7}$$

Using $\hat{\tau} = \theta^q \tau$, it follows from (5.7) that the condition $\widehat{\mathrm{est}}_n \leq \tau$ for an accepted step reduces to $\gamma \leq h^*(1-\theta^q)/\theta^q$, which is our assumption (5.6).

The new stepsize is

$$\begin{aligned} h_{n+1} &= \left(\frac{\hat{\tau}}{\widehat{\mathrm{est}}_{n+1}} \right)^{1/q} h_n \\ &= \left(\frac{h^*}{h^* + \gamma} \right)^{1/q} h_n \\ &= h^{*1/q} (h^* + \gamma)^{1-1/q}. \end{aligned} \tag{5.8}$$

This means that

$$h_{n+1} - h^* = h^* \left(\left(\frac{h^* + \gamma}{h^*} \right)^{1-1/q} - 1 \right). \tag{5.9}$$

We see from (5.8) and (5.9) that

- if $\gamma > 0$, then $h^* < h_{n+1} < h_n$, and
- if $\gamma < 0$, then $h_n < h_{n+1} < h^*$.

In either case, the new stepsize corresponds to a smaller perturbation, with the same sign. In particular, the conditions (5.5) and (5.6) hold on the next step.

With EPUS control, under the assumption (5.5), the error estimate is independent of the stepsize. We have

$$\widehat{\mathrm{est}}_n := |\, \sum_{i=1}^{s} \hat{b}_i f(X_i^*) \,| = \theta^q \tau,$$

and hence the step is accepted and the stepsize remains the same on the next step. ∎

Our approach is now to construct solutions to (5.1), (5.2) and (5.3) for small τ. The theorem above tells us that if h^* and δ can be bounded away from zero and $f_{\max}$ can be bounded above, then the basin of attraction will not shrink to zero as $\tau \to 0$.

The next result, which is essentially an application of the Implicit Function Theorem, shows that a solution for $\tau = 0$ can be extended to a solution for small τ.

Lemma 5.2 *Suppose that the ERK formulas defining Φ and Ψ in (2.1)–(2.2) are distinct with order at least one. Then any S-acceptable solution to (5.1), (5.2) and (5.3) for $\tau = 0$ can be extended to an S-acceptable solution for small τ. Further, the solutions thus generated depend continuously upon τ.*

Proof. Let

$$B = \begin{bmatrix} b_1 & b_2 & \dots & \dots & b_s \\ \widehat{b}_1 & \widehat{b}_2 & \dots & \dots & \widehat{b}_s \end{bmatrix} \in \mathbb{R}^{2\times s}. \tag{5.10}$$

First we prove by contradiction that B has rank two. If $\text{rank}(B) = 1$ then there exists $\alpha \in \mathbb{R}$ such that $b_i = \alpha \widehat{b}_i$ for $1 \le i \le s$. But any ERK formula with order at least one satisfies $\sum_{i=1}^{s} b_i = 1$ (see, for example, [8]). Hence, $1 = \sum_{i=1}^{s} b_i = \alpha \sum_{i=1}^{s} \widehat{b}_i = \alpha$, contradicting the fact that two methods are distinct.

Since B has rank 2, there exist indices i and j, with $i \neq j$, such that

$$\begin{bmatrix} b_i & b_j \\ \widehat{b}_i & \widehat{b}_j \end{bmatrix} \tag{5.11}$$

is nonsingular. Now regard f_k^* for $k \neq i, j$ as fixed. Equations (5.2) and (5.3) become

$$\begin{bmatrix} b_i & b_j \\ \widehat{b}_i & \widehat{b}_j \end{bmatrix} \begin{bmatrix} f_i^* \\ f_j^* \end{bmatrix} = \begin{bmatrix} -\sum_{k\neq i,j} b_k f_k^* \\ \pm\widehat{\tau}/h^* - \sum_{k\neq i,j} \widehat{b}_k f_k^* \end{bmatrix}. \tag{5.12}$$

For definiteness, we will take the plus sign in (5.12). Since the coefficient matrix is nonsingular, (5.12) has a unique solution that is continuous in τ. By assumption, the solution at $\tau = 0$ gives distinct values for $\{X_i^*\}_{i=1}^s$ in (5.1), and hence for sufficiently small τ this distinctness will be preserved. ∎

Next we give conditions under which an S-acceptable solution exists for $\tau = 0$.

Lemma 5.3 *Define B as in (5.10) and let*

$$A = \begin{bmatrix} a_{21} & 0 & \dots & \dots & 0 \\ a_{31} & a_{32} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ a_{s,1} & a_{s,2} & \dots & \dots & a_{s,s-1} \end{bmatrix} \in \mathbb{R}^{(s-1)\times(s-1)}. \tag{5.13}$$

Suppose the system

$$Bx = 0, \tag{5.14}$$

has a solution $x \in \mathbb{R}^s$ such that

1. *$x_1 \neq 0$ and $\{x_i\}_{i=2}^s$ are distinct and nonzero.*
2. *Letting $z = [x_1, x_2, \dots, x_{s-1}]^T$, the product Az has distinct, nonzero elements.*

Then an S-acceptable solution exists for $\tau = 0$.

Proof. We fix $h^* = X_1^* = 1$, and let the solution x define the values $\{f_i^*\}_{i=1}^s$. The constraints (5.2) and (5.3) are then satisfied (for $\tau = 0$). Now the components of $\{X_i^*\}_{i=2}^s$ are given by

$$X_1^* e + h^* A z$$

where $e \in \mathbb{R}^{s-1}$ is the vector of ones. Since Az has distinct, nonzero elements, the set $\{X_i^*\}_{i=1}^s$ is distinct, and the result is proved. ∎

When the number of stages s is greater than two, (5.14) represents an underdetermined system of linear equations. We have seen that for the `ode23.m` pair it is not possible to construct a suitable solution, but usually, when there are more than two stages, there will be many solutions to $Bx = 0$ satisfying the required properties. In particular, if we construct a solution where the X_i^* are not distinct, then we can try perturbing the solution so that the constraints still hold and the $\{X_i^*\}$ are forced apart. It would be cumbersome to give a complete formalisation of this process, and hence in the lemma and theorem below we consider only one choice of perturbation. Despite this restriction, the conditions required in the lemma are likely to hold for most ERK formulas, with the exception of those with the first-same-as-last (FSAL) property. A modified approach that is designed for FSAL formulas is given at the end of the section.

Lemma 5.4 *Suppose the matrix*

$$\begin{bmatrix} b_{s-1} & b_s \\ \widehat{b}_{s-1} & \widehat{b}_s \end{bmatrix}$$

is nonsingular. Let

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} := - \begin{bmatrix} b_{s-1} & b_s \\ \widehat{b}_{s-1} & \widehat{b}_s \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ \widehat{b}_1 \end{bmatrix}. \tag{5.15}$$

Then if the set $\{a_{i,1}\}_{i=2}^{s-1} \cup \{a_{s,1} + a_{s,s-1}\alpha\}$ *contains elements that are all distinct and nonzero, an S-acceptable solution can be constructed for* $\tau = 0$.

Proof. Fix $h^* = 1$. Take any solution to (5.14). (In particular, we could use the solution $x = 0$.) Set $f_i^* = x_i$ and $X_1^* = 1$. If the resulting X_i^* in (5.1) are distinct and $f_1^* \neq 0$ then we are done. If not, then perturb f_i^* to $f_i^* + \epsilon_i^*$ for $i = 1, s-1, s$. For the constraints (5.2) and (5.3) to hold for $\tau = 0$, we require

$$\begin{bmatrix} b_{s-1} & b_s \\ \widehat{b}_{s-1} & \widehat{b}_s \end{bmatrix} \begin{bmatrix} \epsilon_{s-1} \\ \epsilon_s \end{bmatrix} = -\epsilon_1 \begin{bmatrix} b_1 \\ \widehat{b}_1 \end{bmatrix}. \tag{5.16}$$

By construction, from (5.15), we can write

$$\begin{bmatrix} \epsilon_{s-1} \\ \epsilon_s \end{bmatrix} = \epsilon_1 \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{5.17}$$

Now the effect of the perturbations is to change the $\{X_i^*\}$ values in (5.1) according to

$$X_i^* \rightarrow X_i^* + h^* a_{i,1}\epsilon_1, \quad \text{for } 2 \leq i \leq s-1, \tag{5.18}$$

$$X_s^* \rightarrow X_s^* + h^*(a_{s,1}\epsilon_1 + a_{s,s-1}\epsilon_{s-1}). \tag{5.19}$$

From (5.17) we may write (5.19) as

$$X_s^* \rightarrow X_s^* + h^*\epsilon_1(a_{s,1} + a_{s,s-1}\alpha). \tag{5.20}$$

Since $\{a_{i,1}\}_{i=2}^{s-1} \cup \{a_{s,1} + a_{s,s-1}\alpha\}$ are distinct and nonzero, it follows from (5.18) and (5.20) that we can choose ϵ_1 such that $\{X_i^*\}_{i=1}^{s}$ are distinct. (As ϵ_1 is varied, (5.18) and (5.20) describe non-parallel straight lines, and we must avoid points of intersection. Hence, there is only a finite number of unsuitable ϵ_1 values.) ∎

Note that for the `ode23.m` coefficients we have $\alpha = -1$ and so $a_{s,1} + a_{s,s-1}\alpha = 1/4 - 1/4 = 0$, confirming that the technique above cannot be used.

Combining our results leads to the following theorem.

Theorem 5.1 *Suppose that the ERK formulas defining Φ and Ψ in (2.1)–(2.2) are distinct with order at least one. Suppose further that the conditions required for Lemma 5.4 hold.*

Then for any sufficiently small τ, it is possible to construct a locally piecewise constant function f (depending upon τ) such that the ERK algorithm produces a spurious fixed point for all (h_0, y_0) in a nonempty region $\mathcal{B}(\tau)$. Further, the area of $\mathcal{B}(\tau)$ remains $O(1)$ as $\tau \to 0$. The spurious solution has $y_n \equiv y_0$, and the values f_i^* taken by the constant pieces of f converge to finite limits as $\tau \to 0$. Within $\mathcal{B}(\tau)$, for EPS control, $h_n \to 1$ monotonically, and for EPUS control, $h_n \equiv h_0$.

Proof. Lemma 5.4 guarantees that an S-acceptable solution $h^*, \{X_i^*, f_i^*\}$ exists for $\tau = 0$ with $h^* = 1$ and $X_1^* = 1$. Lemma 5.2 then guarantees that an S-acceptable solution exists for sufficiently small τ, with f_i^* varying continuously with τ.

Now, since f_i^* and hence X_i^* in (5.1) depend continuously upon τ, the quantities $f_{\max}$ and δ in (5.4) also depend continuously upon τ. Since δ is bounded away from zero at $\tau = 0$, there exist positive constants $\tau^*, K, k > 0$ such that

$$f_{\max} \le K \quad \text{and} \quad \delta > k, \quad \text{for all } \tau \in (0, \tau^*].$$

It then follows that the constraints (5.5) and (5.6) define a region that does not shrink to zero as $\tau \to 0$. For example, ϵ, γ satisfying

$$|\epsilon| \le k/3, \quad |\gamma| \le \frac{1}{3}\min\{k/(sK\max\{|a_{ij}|\}), (1-\theta^q)/\theta^q\} \tag{5.21}$$

are valid for all $\tau \in (0, \tau^*]$. ∎

We illustrate this theory on the RKF45 pair of Fehlberg [8, page 185]. We use the higher order formula to advance the solution, with EPS control and with $q = 5$ and $\theta = .8$ in (1.8). This agrees with the implementation of RKF45 in Matlab's `ode45.m` function [9]. The conditions of Lemma 5.4 hold, and choosing $x = 0$ and $\epsilon_1 = 20$ leads to well-separated X_i^* values. We can use $i = s-1$ and $j = s$ in Lemma 5.2, and taking $\tau = 10^{-10}$ for the tolerance gives the values

$$X^* = \begin{bmatrix} 1 \\ 6 \\ 2.88 \\ 18.6 \\ 41.6 \\ -8.1 \end{bmatrix}, \quad f^* = \begin{bmatrix} 20 \\ 0 \\ 0 \\ 0 \\ 11.57407407391024 \\ -7.893518519329516 \end{bmatrix}.$$

(The fixed point is not sensitive to small changes in X_i^*, hence we do not to give these values to high precision.) It follows that we can take $I_i = [X_i^* - \delta, X_i^* + \delta]$, with $\delta = 1/2$ and $f_{\max} = 20$ in (5.4). The RKF45 pair has $s = 6$ and $\max\{|a_{ij}|\} = 8$. Hence, applying Lemma 5.1, it follows that $h_0 = 1 + \gamma$ and $y_0 = 1 + \epsilon$ will lead to an iteration for which $y_n \equiv y_0$ (spurious) and h_n converges monotonically to 1, whenever

$$|\gamma| < \frac{1}{2 \times 6 \times 20 \times 8} = \frac{1}{1920}, \quad |\epsilon| < \frac{1}{2} - (6 \times 20 \times 8)|\gamma| = \frac{1}{2} - 960|\gamma|. \tag{5.22}$$

This has been confirmed numerically.

The conditions required by Lemma 5.4 do not hold for the class of so-called first-same-as-last (FSAL) ERK pairs, but only a minor change is required to make the technique applicable. Since FSAL pairs are widely used, we give the details here.

Definition: An ERK pair (1.2)–(1.4) is said to have the *FSAL property* if

$$b_j = a_{s,j} \quad \text{for} \quad 1 \le j \le s-1 \qquad \text{and} \quad b_s = 0. \tag{5.23}$$

The FSAL property offers the practical advantage that the first stage k_1 on one step is identical to the last stage k_s on the previous step. Hence, the expense of one evaluation of the function f can be avoided. If the FSAL conditions (5.23) hold then $a_{s,1} + a_{s,s-1}\alpha = 0$ in Lemma 5.4, and hence the technique breaks down. This can also be seen from the fact that at a fixed point, (5.2) gives

$$X_s^* = X_1^* + h^* \sum_{j=1}^{s-1} a_{sj} f_j^* = X_1^* + h^* \sum_{j=1}^{s-1} b_j f_j^* = X_1^*,$$

and hence, for a FSAL pair, the $\{X_j^*\}$ can never be distinct. The technique used in Lemma 5.4 is, however, easily adapted to allow for this, as we now show.

Lemma 5.5 *Given an ERK pair with the FSAL property, suppose the matrix*

$$\begin{bmatrix} b_{s-2} & b_{s-1} \\ \widehat{b}_{s-2} & \widehat{b}_{s-1} \end{bmatrix}$$

is nonsingular. Let

$$\begin{bmatrix} \lambda \\ \rho \end{bmatrix} := -\begin{bmatrix} b_{s-2} & b_{s-1} \\ \widehat{b}_{s-2} & \widehat{b}_{s-1} \end{bmatrix}^{-1} \begin{bmatrix} b_1 \\ \widehat{b}_1 + \widehat{b}_s \end{bmatrix}. \tag{5.24}$$

Then if the set $\{a_{i,1}\}_{i=2}^{s-2} \cup \{a_{s-1,1} + a_{s-1,s-2}\lambda\}$ *contains elements that are all distinct and nonzero, a solution to (5.1), (5.2) and (5.3) for* $\tau = 0$ *can be constructed with* $\{X_i^*\}_{i=2}^{s}$ *distinct,* $X_1^* = X_s^*$ *and* $0 \ne f_1^* = f_s^*$.

Proof. The proof is similar to that of Lemma 5.4. Since $X_1^* \equiv X_s^*$ we must keep $f_1^* \equiv f_s^*$. Hence, we make an extra perturbation (to f_{s-2}).

Fix $h^* = X_1^* = 1$, and set $f_i^* = 0$ for $1 \le i \le s$. Now perturb

$$\begin{aligned} f_1^* &\to f_1^* + \epsilon_1, \\ f_s^* &\to f_s^* + \epsilon_1, \\ f_{s-2}^* &\to f_{s-2}^* + \epsilon_{s-2}, \\ f_{s-1}^* &\to f_{s-1}^* + \epsilon_{s-1}. \end{aligned}$$

For the constraints (5.2) and (5.3) to hold for $\tau = 0$, we require

$$\begin{bmatrix} b_{s-2} & b_{s-1} \\ \widehat{b}_{s-2} & \widehat{b}_{s-1} \end{bmatrix} \begin{bmatrix} \epsilon_{s-2} \\ \epsilon_{s-2} \end{bmatrix} = -\epsilon_1 \begin{bmatrix} b_1 \\ \widehat{b}_1 + \widehat{b}_s \end{bmatrix},$$

which, by construction, is

$$\begin{bmatrix} \epsilon_{s-2} \\ \epsilon_{s-1} \end{bmatrix} = \epsilon_1 \begin{bmatrix} \lambda \\ \rho \end{bmatrix}.$$

The effect of the perturbations is to change the $\{X_i^*\}$ values according to

$$\begin{aligned} X_i^* &\rightarrow X_i^* + h^* a_{i,1}\epsilon_1, \quad \text{for } 2 \le i \le s-2, \\ X_{s-1}^* &\rightarrow X_{s-1}^* + h^*\epsilon_1(a_{s-1,1} + a_{s-1,s-2}\lambda), \end{aligned}$$

with X_1^* and X_s^* unchanged.

Since $\{a_{i,1}\}_{i=2}^{s-2} \cup \{a_{s-1,1} + a_{s-1,s-2}\lambda\}$ are distinct and nonzero, it follows that we can choose ϵ_1 such that $\{X_i^*\}_{i=2}^s$ are distinct, as required. ∎

The following theorem then holds.

Theorem 5.2 *Suppose that the ERK formulas defining Φ and Ψ in (2.1)–(2.2) are distinct with order at least one, and have the FSAL property, Suppose further that the conditions required for Lemma 5.5 hold.*

Then for any sufficiently small τ, it is possible to construct a locally piecewise constant function f (depending upon τ) such that the ERK algorithm produces a spurious fixed point for all (h_0, y_0) in a nonempty region $\mathcal{B}(\tau)$. Further, the area of $\mathcal{B}(\tau)$ remains $O(1)$ as $\tau \rightarrow 0$. The spurious solution has $y_n \equiv y_0$, and the values f_i^* taken by the constant pieces of f converge to finite limits as $\tau \rightarrow 0$. Within $\mathcal{B}(\tau)$, for EPS control, $h_n \rightarrow 1$ monotonically, and for EPUS control, h_n remains constant for all n.

Proof. Lemma 5.5 produces a solution for which $\{X_i^*\}_{i=2}^s$ are distinct, $X_1^* = X_s^*$ and $0 \neq f_1^* = f_s^*$. Hence, it is possible to fit a piecewise constant function through this data.

By assumption, with $i = s-2$ and $j = s-1$ the matrix (5.11) is nonsingular. Hence, for any small τ, solving (5.12) produces a solution for which f_1^* and f_s^* are unchanged and $\{X_i^*\}_{i=2}^s$ remain distinct.

Now, using the approach in Lemma 5.1 completes the proof. ∎

This technique has been used to create spurious solutions with the DOPRI(5,4) pair [8, page 186] of Dormand and Prince.

6 Conclusions

Solving an initial value ODE system is often part of a more complicated problem, and an adaptive ODE solver is frequently used as "black-box" whose output is fed, unmonitored, into another algorithm. For this reason it is important to know what guarantees can be made about the numerical solution. In this work we have examined the potential for spurious fixed points. The results are couched in terms of the accuracy parameter τ used in the ODE solver, and they apply to all widely-used local error control and stepsize selection algorithms for explicit Runge-Kutta formulas.

Our main result is positive. When standard local error control is used, the chance of encountering spuriousity is extremely small. For general systems of ODEs, the constraints imposed by the *error control criterion* make spuriosity extremely unlikely. For scalar problems however, the mechanism by which the algorithm succeeds is indirect—spurious fixed points are not removed, but those that exist are forced by the *stepsize selection mechanism* to be locally repelling (with the relevant eigenvalues behaving like $O(1/\tau)$). More precisely, there is a hierarchy of unfortunate behaviour:

1. The adaptive method admits a spurious fixed point.

2. The adaptive method admits a stable, spurious fixed point.

3. The adaptive method admits a stable, spurious fixed point with a significant basin of attraction.

In section 2, we argued that level 1 behaviour is not uncommon in the scalar case. Spurious fixed points arise generically whenever the fixed-stepsize spurious fixed point branches of the individual formulas intersect. However, level 2 is unlikely to arise—Theorem 3.1 shows that, for small τ, all but a pathological class of cases can be ruled out. In particular, only branches that intersect *tangentially* can produce stable fixed points. Even when a stable, spurious fixed point exists, the eigenvalues of the Jacobian are likely to become large away from a small neighbourhood of the fixed point. Hence, level 3 behaviour, which is the most practically significant scenario, is extremely unlikely.

It is possible, though, in general, to construct ODEs for which an adaptive ERK method will behave badly. Sections 4 and 5 show how this can be done using a locally piecewise constant function f in (1.1). Since the disjoint pieces can be connected in any manner, f can be made arbitrarily smooth. Hence, smoothness of f alone is not sufficient to guarantee that spurious behaviour will be eliminated. These examples highlight the worst-case behaviour of adaptive ERK methods.

Acknowledgement
The work of the three authors was supported by grant GR/H94634 from the Engineering and Physical Sciences Research Council (formerly the Science and Engineering Research Council).

References

[1] D.F. Griffiths, P.K. Sweby, and H.C. Yee, *On spurious asymptotic numerical solutions of explicit Runge-Kutta methods*, IMA J. Numer. Anal., 12 (1992), pp. 319–338.

[2] E. Hairer, A. Iserles, and J.M. Sanz-Serna, *Equilibria of Runge-Kutta methods*, Numerische Mathematik, 58 (1990), pp. 243–254.

[3] G. Hall, *Equilibrium states of Runge-Kutta schemes*, ACM Transactions on Mathematical Software, 11 (1985), pp. 289–301.

[4] G. Hall and D.J. Higham, *Analysis of stepsize selection schemes for Runge-Kutta codes*, IMA J. Numer. Anal., 8 (1988), pp. 305–310.

[5] D.J. Higham and A.M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual*, tech. report, Stanford University (in preparation), 1994.

[6] A.R. Humphries, *Spurious solutions of numerical methods for initial value problems*, IMA J. Numer. Anal., 13 (1993), pp. 263–290.

[7] A. Iserles, *Stability and dynamics of numerical methods for nonlinear ordinary differential equations*, IMA J. Numer. Anal., 10 (1990), pp. 1–30.

[8] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems*, Wiley, 1991.

[9] The MathWorks, Inc., *MATLAB User's Guide*, Natick, Massachusetts, 1992.

[10] J.M. Sanz-Serna, *Numerical ordinary differential equations vs. dynamical systems*, in The Dynamics of Numerics and the Numerics of Dynamics, D.S. Broomhead and A. Iserles, eds., The Institute of Mathematics & its Applications, 1992, pp. 81–106.

[11] L.F. Shampine and R.C. Allen, *Numerical Computing: An Introduction*, Saunders, Philadelphia, 1973.

[12] A.M. Stuart and A.R. Humphries, *The essential stability of local error control for dynamical systems*, SIAM J. Numer. Anal., To appear.

[13] H.C. Yee, P.K. Sweby, and D.F. Griffiths, *Dynamical systems approach study of spurious steady state numerical solutions of nonlinear differential equations. 1. The ODE connection and its implications for algorithm developments in computational fluid dynamics*, J. Comp. Phys., 97 (1991), pp. 249–310.